人の行動目的理解のための動画像からの物体-行動間の関連付けの検証

○横田 雅恵（中央大学），Pathak Sarthak（中央大学）

新妻 実保子（中央大学），梅田 和昇（中央大学）

○ Masae YOKOTA (Chuo University), Sarthak PATHAK (Chuo University),

Mihoko NIITSUMA (Chuo University), and Kazunori UMEDA (Chuo University)

Abstract: We aim to recognize human behavior from image information and to infer a person's desired goal from a series of actions. In this paper, we have conducted two types of validation via LLM to infer the purpose of an action from the history of multiple actions, and one type of validation via VLM to infer the purpose of an action from an image. Through these verifications, we confirmed that LLM can select the most likely purpose of action from sentences in the history of actions, and that VLM can roughly infer the purpose of a person's action from images. Based on the verification results, we proposed a framework for inferring the purpose of action.

1. 緒言

近年，ロボットと人の複数の協働作業において人にかかる負荷を減らすために，柔軟なヒューマンロボットコミュニケーション手法が研究されている．協働作業において，人がロボットに対して作業ごとに指示や操作を行うと人の作業負荷は新たに増えてしまう．そのため，ロボットが能動的に動くことで，人間同士でのような自然な連携による効率化を達成することが望ましい[1]．人からの明示的な指示をただ待つのではなく，人の動きからロボットが動作の文脈を理解し，自身の役割を理解し先回りすることで人の作業負担を軽減することが出来る．ここで，ロボットが情報を取得し自身の役割を解釈する課題に対して，様々なアプローチで研究が行われている[2]．

本論文では，**Fig. 1** に示すような行為の 7 段階モデル[3] に基づいて，人の一連の行動には明確な意図が潜在していると考える．例えば，買い物に行こうという意図が形成されると，人は “鞄を手に取る”，“財布を鞄に入れる”，“上着を着る”，“玄関へ向かう”，“靴を履く”，“ドアを開ける” という詳細な行為を計画し，実行していく．この時，実行された複数の行為の履歴から人が抱く意図を文脈的に推測することで，推定された時点以降の行為について予測することができる．本研究の目標は，この予測結果に基づいてロボットが先回りして作業を支援することである．

本論文では認識された行為の履歴から潜在する意図を推定するために大規模言語モデル（LLM）[4] を用いて検証を行った．まず，3.2 では物体と行動がセットになった行為の履歴から人の達成したい目標を推測し文章として生成できることを確認し，複数モデル間で結果を比較した．次に，3.3 では行為の履歴から推測される人の意図を，あらかじめ用意した意図のリストの中から最も可能性の高いものを選び，結果を複数モデル間で比較した．これらの検証から，複数の行動から人の動作の意図を推測するために用いる LLM モデルを決定するために重要である点を明らかにした．また，3.4 では，動画像に対し大規模視覚言語モデル（VLM）[5] を用い，物体情報をプロンプトに追加することで動画像中の人物の行動の意図を推測する検証を行った．これらの検証を通して，下記の 3 点を明らかにし，動作意図推定のためのフレームワークを提案する．

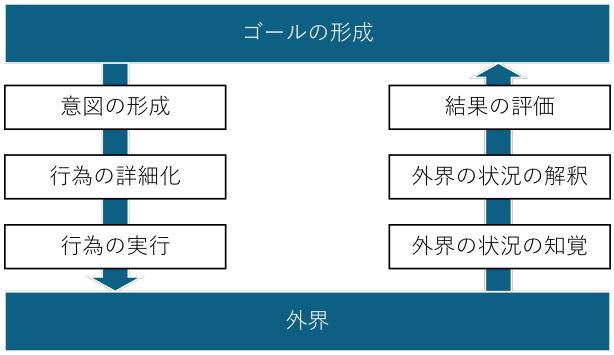


Fig. 1 Norman's seven-stage theory of action

- 行動履歴の文章から，人の行動目的を推測する文章の生成結果がモデルの性能に大きく依存すること．
- 行動履歴の文章から，人の行動目的リスト内から最も類似度の高い単語を選ぶ結果はモデルの性能に依存しにくいこと．
- VLM により動画像から大まかに人の行動目的を推測することができること．

2. 関連研究

人とロボットが協働する際に，画像情報を用いて自然かつ直感的なコミュニケーション方法を構築することで協働の有効性を向上させることを狙って，研究が行われている[6]．

これらの手法の中では，人の行動を観測することで人の動きの意図を理解し，ロボットが積極的に行動して人の作業を助ける研究がされている[7]．ここで，人の行動の観察結果から人の動きの意図の理解を行うために，時系列や文脈を考慮することが重要である．しかし，一つ一つの行動やアクセスしている物体ごとに行動の意味を説明するモデルを一から構築するためには膨大なデータセットや時間が必要となるため，困難である．そこで，本論文では常識的な知識が事前に学習済みであり，文脈を理解できるとされている LLM を用いることで，人の一連の行動と物体との関係から行動の目的を推定できる可能性があると考えた．多くの場合で生成した目的と，実際の目的とが一致することを期待して，検証を行った．

また，人と物のインタラクションを考慮した説明文

Table 1 Result of the first valification

モデル	生成結果
GPT-2	結果なし
Llama2	“relax” という語を使った文章

を画像から生成する研究も行われている[8]．そこで，本論文では動画像から人の周囲の物体に着目し，人が達成できる意図リストを事前に用意し，物体検出アルゴリズムと VLM により最も互換性の高い意図を推定した．

以上より，本論文では下記のような検証を行い，人の行動から動作意図を推定するフレームワークを構築するために考察を行った．

- 検証 1 では，行動の履歴の文章から人の行動の目的を推定する文章を生成する手法を試した．
- 検証 2 では，行動の履歴の文章から推測される行動の目的を，予め用意した意図のリストの中から最も可能性の高い意図を選ぶ手法を試した．
- 検証 3 では，動画像中の人の周囲の物体情報より，予め用意した意図のリストの中から最も可能性の高い意図を選ぶ手法を試した．

3. 検証

3.1 手法概要

本論文では，行動の履歴から動作意図を推定する予備実験として，3 種の検証を行った．検証 1 と検証 2 は LLM を用いて人の行動履歴から行動の目的を推定する検証であるが，検証 1 では目的を説明する文を生成する手法を試したことに対し，検証 2 では予め用意した行動目的の単語のリストから最も可能性の高いものを選ぶ手法を試した．また，検証 3 では VLM を用いて，動画像中の人の周囲の物体情報から，予め用意したリスト中で最も可能性の高い意図を選ぶ手法を試した．検証 1 と検証 2 では GPT-2[9] と Llama2[10] を用いた．検証 3 では YOLOv8[11] と CLIP[12] を用いた．

3.2 検証 1：LLM による行動目的文の生成

“何も持たずに歩いている”，“本を持っている”，“カップを持っている”，“本とカップを持って歩いている”，“椅子に座っている”，“カップで飲み物を飲んでいる”，“本を読んでいる” という行動履歴を用意した．これらの行動履歴から人の行動目的を推測した文を 3 回ずつ生成させたところ，結果は **Table 1** のようになった．GPT-2 では入力したプロンプトの最後の一文を繰り返し出力した結果となり，Llama2 では人がリラックスしようとしていると考えられる旨の文章が出力され，期待していた結果が得られた．

3.3 検証 2：LLM による行動目的の選択

検証 2 では，検証 1 で用いた行動履歴に加え，“take a break”，“study”，“sleep” という行動目的のリストを用意し，行動履歴から最も可能性の高いと考えられる行動目的を選ぶ手法を試したところ，結果は **Table 2** のようになった．GPT-2 ではリスト内の語の中から “take a break” を選ばれ，期待していた結果となった．しかし，Llama2 ではリスト内にない語が追加され，その中から “relax” が選ばれた．

Table 2 Result of the second valification

モデル	選択結果
GPT-2	take a break (リスト内から選出)
Llama2	relax (リスト外の語が追加された)

Table 3 Result of the third valification

行動目的	平均類似度
take a break	0.48
work	0.31
cook	0.02
sleep	0.05

3.4 検証 3：VLM による動画像からの意図推測

本検証では，YOLOv8 を用いて物体検出を行い，検出された物体名を CLIP のプロンプトに追加するという手法で，予め用意した行動目的のリストの内，画像の解釈としての類似度を各フレームで計算した．動画は 503 枚のフレームであり，人が本を持って机と椅子に近づき，着席ののち，卓上のカップから飲み物を飲み，読書をするという内容のものである．結果は **Table 3**，**Fig. 2** のようになった．この時，行動目的のリストは “take a break”，“work”，“cook”，“sleep” の 4 種であり，“take a break” が最も高い類似度になることを期待した．

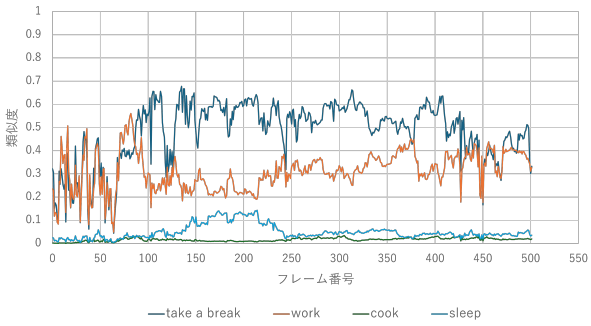
Table 3 より “take a break” が最も類似度が高くなることを確認できた．また，**Fig. 2(a)** はフレームごとの類似度を行動目的別にグラフにプロットしたものであり，**Fig. 2(b)** はそのフレーム時点での平均類似度を行動目的別に表したものである．Fig. 2(a) では初めのフレーム時点では “take a break” と “work” の類似度が近いが，Table 3 と Fig. 2(b) では，期待していた結果である，“take a break” の行動目的が最も類似度が高いという結果が得られた．

4. 考察

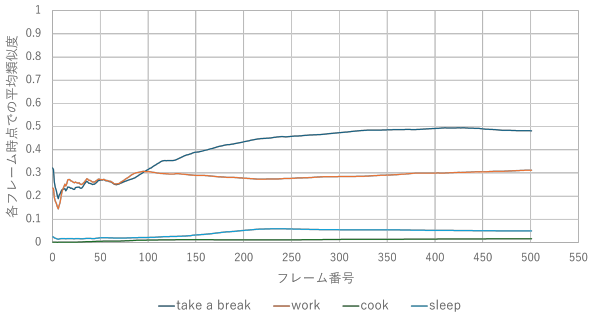
4.1 検証 1 と検証 2 の考察

検証 1 と検証 2 では LLM を用い，行動の履歴を含んだプロンプトによる行動目的の推測を行った．検証 1 ではプロンプトに対して一から回答文を生成する手法を試した．Table 1 より，モデル間で大きな違いが出ることが分かった．Llama2 のほうが GPT-2 よりも優れているとされているため，期待していた結果が得られたと考えられる．一方で，一から文章を生成する手法では回答が不安定になる可能性があり，実用化に向けた評価が難しい．そこで，検証 2 では予め用意した行動目的のリスト内から最も関連性の高い語を選ぶ手法を試した．Table 2 より，GPT-2 では期待していた通りの結果が得られたが，Llama2 ではリストにない語を追加し，それを選ぶという結果となった．しかし，選ばれた語と元々リスト内にあった語と意味が近いため，文脈理解としては間違った結果ではないと考えられる．

これらの結果から，LLM により行動履歴から行動目的を推測する方針において，一から文章を生成する手法ではなく，予め用意した行動目的のリスト内から選ぶ手法のほうがより確実であると考えられる．しかし，Llama2 では類似語が勝手に追加され選ばれていたた



(a) Frame-by-frame similarity



(b) Average similarity at that frame

Fig. 2 Similarity graph of the third valification

め，改善が必要である．また，今回は，今までの文脈のみから次に来る語や文章を推論するモデルのみを使用したため，不安定な結果になったと考えられる．今後は，類似語での推論結果の扱いについての対処を考慮するとともに，BERT[13] のような前後の文脈を踏まえて推論を行うモデルも使用し，結果を比較する必要がある．

4.2 検証 3 の考察

検証 3 では VLM を用い，動画像から人の行動目的を推測する手法を試した．物体検出アルゴリズムを組み合わせ，人の周囲の物体名をプロンプトに追加し，そのフレームでの画像に対して人の行動目的を推測し，予め用意した行動目的のリスト内の語と各フレーム時点での画像との平均類似度を算出した．Table 3，Fig. 2 より，期待していた通りの結果が得られた．

これらの結果から，VLM を用いて大まかに行動目的を推測できることが分かった．一方で，画像間の時系列を加味した推測結果ではないため，動画像の内容を決定することはできない．今後は，画像間の時系列を考慮し，行動目的の推測を向上させる手法を検討する．

4.3 動作意図推定フレームワークの提案

4.1 より，LLM による行動履歴から文脈的に行動目的を推測することが出来ると分かった．また，4.2 では VLM により画像から大まかに人の行動目的を推測できると分かった．これらの結果から，**Fig. 3** に示すような動作意図推定フレームワークを提案する．

まず，多くの行動目的を含むリストを作成する．次に，物体検出アルゴリズムと VLM により，その環境下で達成でき，かつ画像から推測される行動目的をある程度まで絞り込み，新たな行動目的リストを作成する．例えば，調理器具が検出できない環境では行動目

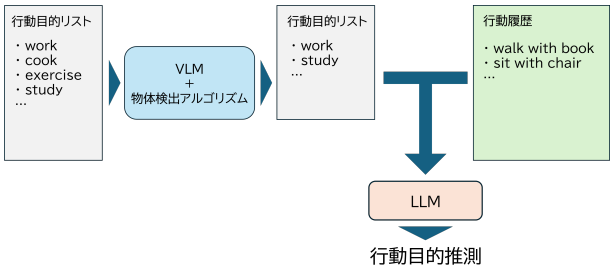


Fig. 3 Proposed method

的リスト内から “cook” が除外される．続いて，姿勢認識アルゴリズムと物体検出アルゴリズムを組み合わせて “action with object” という形式で行動を説明する文章を生成する．この説明文を蓄積した行動履歴から LLM による行動目的推測を行う．

5. 結言

本論文では，人とロボットの共同作業においての柔軟かつ負荷の少ないヒューマンロボットコミュニケーション手法として行動目的推測のフレームワーク構築を目指し，LLM や VLM を用いた検証を行った．

検証 1 と検証 2 では，LLM を用いて人の複数の行動から達成したい目的を文脈的に推測するために 2 種の手法を試し，複数の行動説明文から行動目的が推測されたことを確認した．また，検証 3 では，物体検出アルゴリズムと VLM を用いて動画像から人の行動目的を推測し，フレーム時点での平均類似度の結果から，大まかな行動目的推測を行えることを確認した．最後に，これらの検証を通して，Fig. 3 に示すような動作意図推定のフレームワークを提案した．

今後は，下記の 3 点に留意し，提案したフレームワークの実現を進める．

- LLM の他のモデルとの比較を行い，適切なモデルを選択する．
- LLM の利用において類似語の仕様による推測結果の変化を観察し，プロンプトエンジニアリングを行う．
- VLM の利用において，画像間での時系列を考慮し，行動目的の推測を向上させる．

参考文献

[1] D. Mukherjee et al. A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings. Robotics and Computer-Integrated Manufacturing 73, (2022).

[2] Y. Zhang and T. Doyle. Integrating intention-based systems in human-robot interaction: a scoping review of sensors, algorithms, and trust. Frontiers in Robotics and AI 10, (2023).

[3] D. A. Norman. 誰のためのデザイン? : 認知科学者のデザイン原論. 新潮社, (1990).

[4] F. Yu et al. Natural Language Reasoning, A Survey, (2023). arXiv: 2303.14725 [cs.CL]. URL: https://arxiv.org/abs/2303.14725.

[5] J. Zhang et al. Vision-Language Models for Vision Tasks: A Survey, (2024). arXiv: 2304.00685. URL: https://arxiv.org/abs/2304.00685.

[6] N. Robinson et al. Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic

Review. ACM Transactions on Human-Robot Interaction 12.1, pp. 1–66, (2023). URL: http://dx.doi.org/10.1145/3570731.

[7] E. V. Mascaro, D. Sliwowski, and D. Lee. HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs. (2024). arXiv: 2309.16524. URL: https://arxiv.org/abs/2309.16524.

[8] H. Zhang et al. Enhancing Human-Centered Dynamic Scene Understanding via Multiple LLMs Collaborated Reasoning, (2024). arXiv: 2403.10107. URL: https://arxiv.org/abs/2403.10107.

[9] J. Radford A.and Wu et al. Language Models are Unsupervised Multitask Learners, (2019).

[10] H. Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023). arXiv: 2307.09288. URL: https://arxiv.org/abs/2307.09288.

[11] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction 5.4, pp. 1680–1716, (2023). DOI: 10.3390/make5040083. URL: https://www.mdpi.com/2504-4990/5/4/83.

[12] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[13] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019). arXiv: 1810.04805. URL: https://arxiv.org/abs/1810.04805.